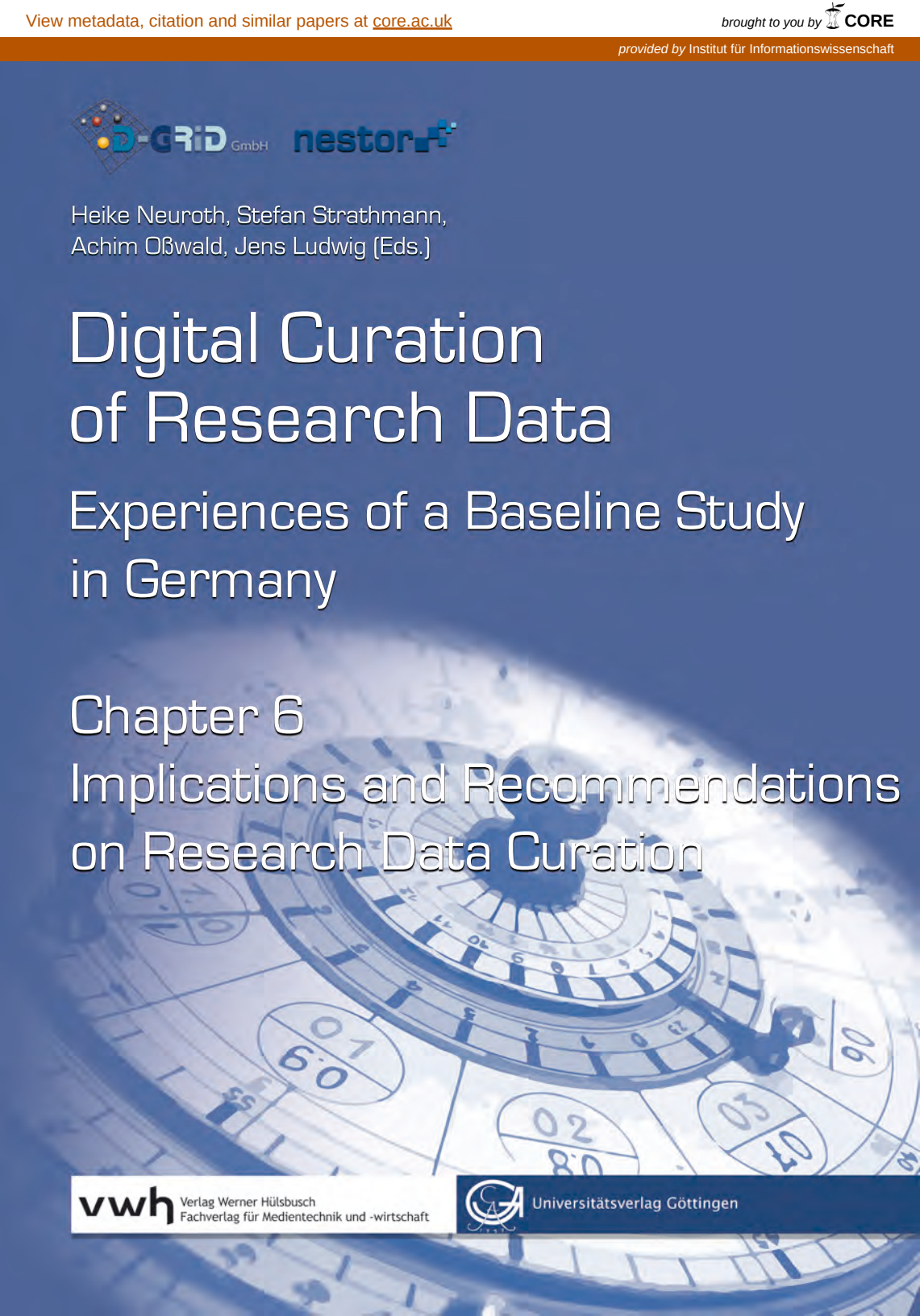
Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

Chapter 6
Implications and Recommendations on Research Data Curation



vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

**Heike Neuroth, Stefan Strathmann,
Achim Oßwald, Jens Ludwig (Eds.)**

Digital Curation of Research Data

Experiences of a Baseline Study in Germany

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Digital Curation of Research Data

Herausgegeben von Heike Neuroth, Stefan Strathmann, Achim Oßwald und Jens Ludwig · im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland · http://www.langzeitarchivierung.de/

Edited by Heike Neuroth, Stefan Strathmann, Achim Oßwald and Jens Ludwig · within the context of nestor – Network of Expertise in the Long-Term Storage of Digital Resources for Germany · http://www.langzeitarchivierung.de/

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de abrufbar.

Bibliographic information of the German National Library
The German National Library lists this publication in the German National Bibliography; detailed bibliographic data is available online at http://www.d-nb.de.

Die Inhalte dieses Buches stehen auch als Onlineversion über die Website von nestor zur Verfügung / This work is available as an Open Access version at the nestor website: http://nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php?lang=en

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

in Kooperation mit dem Universitätsverlag Göttingen
in cooperation with the Universitätsverlag Göttingen

Printed in Poland · ISBN: 978-3-86488-054-4

Content

6 Implications and Recommendations on Research Data Curation

Heike Neuroth, Achim Oßwald, and Uwe Schwiegelshohn

On the basis of the comparative survey of approaches to the management of research data in the eleven academic disciplines that have been analysed in the course of this survey, the following results and theses can be formulated. They emphasize the importance of research data curation from a research perspective and refer to conceptual and operational circumstances that should be considered as an initial result and looked at more closely. However, a number of aspects in terms of science and social policy have to be considered as well.

General Issues:

- The importance of research data and its long-term storing and provision is emphasized by all academic disciplines surveyed here.
- The different approaches to research data curation research data curation in these disciplines do not indicate a lack of cooperation across disciplinary boundaries but are a logical consequence of the different requirements and methods practiced within every single discipline.
- Cooperative structures within a discipline are the rule rather than an exception in the field of research data curation.
- Infrastructure facilities such as libraries or data centers are often included as cooperation partners in research data curation. However, their role and current function has not been clearly defined yet.
- In many academic disciplines, researchers are still confronted with a lack of appreciation for the value of long-term archiving and a low acceptance for data sharing and the re-use of data. The awareness of academic disciplines as well as of society and other stakeholders (e.g. libraries, data centers etc.) for the value of data is an important precondition for further discussions and developments.
- Data management, one of the first steps of the actual research data curation, comprises subject-specific as well as generic tasks. The

close cooperation of the various interest groups and stakeholders allows the exact definition of tasks and areas of responsibility.

- It is not possible to make any reliable statements about the data volumes and the number of digital objects that are to be stored and provided, neither for single academic disciplines nor for disciplines in general. All in all, however, a rapid increase of the data volume of digital research data can be recognized across all academic disciplines.

Research Data Centers:

- The processes for ensuring research data curation are already better established in those disciplines in which central structures for data management have emerged than in other academic disciplines.
- Many disciplines consider data centers as an ideal solution for improving and securing the availability and the efficient re-use of research data in the long term. They can be organized centrally or in a decentralized network. They may also play an important role in the development of standards and in providing advice within the relevant academic disciplines.
- There is a need for clarification regarding the reliability of data centers and what criteria have to be met to ensure it. Questions about how to evaluate a data center's trustworthiness (e.g. through the external certification of data centers) and who is responsible for doing so are still open.

Metadata and Formats:

- Nearly every academic discipline uses its own metadata formats. Most of these formats are based on XML. Many academic disciplines have developed subject-specific metadata formats in recent years.
- Research data are available in an almost unmanageable number and variety of data formats. Almost all disciplines share the common trait of using numerous subject-specific and proprietary formats.
- The individual disciplines handle the diversity and heterogeneity of formats very differently. The different formats are either specified by a policy or otherwise restricted, or the choice of format is open or rather cannot be limited because of discipline-related reasons.

- Overall, academic disciplines use open formats wherever they can. However, this can be severely restricted by the given software or hardware. Considering established industrial and commercial processes is helpful when implementing standardization.

Technical Backup of Data:

- The technical backup of data is a first step to research data curation. Through the purely technical storage of research data, the integrity of the data can be preserved, independent of file and metadata formats. However, this does not guarantee the effective re-use of research data.
- Limiting the variety of data and metadata formats reduces the number of technical environments (regarding both hardware and software) necessary for reproducing the data. This makes its re-use easier.
- To ensure the technical and intellectual re-usability of research data, continuous technology watch, the observation of requirements and technical equipment and community watch are needed.

Re-use of Research Data:

- Research data are made available for re-use for various reasons, e.g. for cooperation within research projects, for external researchers or for the general (professional) public upon publication.
- Academic disciplines, their funding agencies and the general public are following the debates about the re-use of research data and the regulations concerning it. The results are insistent requests to make research data accessible and to guarantee their subsequent re-use in the long term.
- Providing and thus encouraging the re-use of research data is mostly prevented for the following reasons: the threat of loss of control over research data, unsolved legal rights issues and conditions of use concerning data, and data protection restrictions. Potential scenarios for re-use are also influenced by the financial effort involved in generating the data.
- The possibility of long-term citation and referencing of research data is one of several motives for research data curation. Therefore, persistent identifiers play an important role.

Costs, Financing, Efficiency and Institutionalization:

- Since research in general is a responsibility of society as a whole, the cost of research should be paid by public funding. In return, society has the right to expect an efficient use of the resources provided. Regarding research data and their subsequent re-use, there are two approaches:
  - Preserving research data after its creation for re-use, or
  - Reproducing or recreating research data. It must be noted that some processes cannot be repeated, e.g. the collection of climate data.
- If both approaches are roughly equal regarding the quality of research data itself, the most cost-efficient approach is to be preferred. For evaluating these approaches and decision-making, very well-informed cost estimates must be available.
- Currently, there are only limited amounts of reliable information about the costs and cost factors involved in research data curation available. In this respect, it is not possible to make any specific statements about cost structures yet. Previous studies indicate that staff costs represent the largest part of the total costs. Up to now, this staff has been paid mainly from project funds.
- (Proportional) financial coverage for research data curation in the form of institutionally-based funding could only be established in some of the academic disciplines studied here. Most disciplines are still using project funds to finance these activities, although some of these projects have extremely long terms.
- There is an urgent need to clarify the costs and cost factors that arise in the context of research data curation. This is the only way to develop and implement sustainable organizational and business models (including financing models) in the different disciplines.
- Securing and maintaining research data is part of scholarly work. Necessary resources for this must be included in cost estimates for research projects.
- The founding of data centers can help to increase the effectiveness of research data curation. This can lead to new organizational structures that extend beyond institutional boundaries.

Qualification:

- There is an urgent need for training in the field of research data curation, especially in theoretical and conceptual areas. Apart from the nestor activities, there are currently few or no systematic training opportunities available in Germany, neither for researchers on a disciplinary level nor for information specialists.
- The exchange of research results relevant to research data curation or examples of best practices occurs only on a limited basis owing to a lack of systematic opportunities for transferring knowledge. Case-by-case decision-making and the focus on one specific discipline and its perceived uniqueness rather than on shared interdisciplinary characterristics have been hindering the establishment of multidisciplinary evaluation criteria and training measures.
- The integration of research data curation into the methodological principles of degree programs or major study courses (such as data librarian and data curator) and research contexts should be a long-term objective. Additional educational opportunities, such as core subjects or interdisciplinary master's programs, are an important source of support for infrastructure activities.

Social Significance:

- Research results are increasingly driving socio-political decisions concerning issues like nuclear power, pre-implantation diagnostics (PID), pandemics, and health risks. It is necessary to preserve the research data on which these decisions are based so that full transparency will be possible in future evaluations.
- The preservation of our cultural heritage is recognized as a social responsibility. Research data are part of this cultural heritage.
- Investigations into violations of scholarly best practices or the detection of methodological errors require that those research data which formed the basis of the specific publication or research paper continue to be available.

All in all, the results and theses presented in this survey demonstrate how significant the (future) role of research data curation is. The EU expert

group "High Level Expert Group on Scientific Data"[137] gave a similar statement: data are infrastructure and serve as a guarantee for innovative research.

Recommendations for future activities that address this topic must be developed on science policy level and implemented through political and funding programs on a national and international level. Some areas of activity have already been identified by researchers and policy-makers.

The above-mentioned "High Level Expert Group on Scientific Data" gave six recommendations[138] which include the establishment of an international framework for the development of collaborative data infrastructures, increased funding for the development of data infrastructures and the development of new approaches and methods to measure and evaluate the value, the importance and the quality of data use. In addition, the importance of training a new generation of "data scientists" is emphasized as well as the establishment of educational opportunities in the new degree programs. The creation of incentive systems in the field of "green technologies" to meet the increasing demand for resources such as energy plays an important role under environmental aspects as well. Lastly, the recommendations suggest the establishment of an international expert panel to promote and manage the development of data infrastructures.

The report from the Commission on the Future of Information Infrastructure (Kommission Zukunft der Informationsinfrastruktur [KII])[139] emphasizes from a national perspective the need for data management plans and data policies as a prerequisite for the exchange and re-use of research data. These plans and policies need to include clear definitions of the responsibilities, the functions and the roles of all stakeholders. Additionally, specific funding programs are recommended for the various aspects in the field of research data curation, making a distinction between development costs for the construction or upgrading of infrastructure and operating costs for permanent operation, including data maintenance.

137 High Level Expert Group on Scientific Data (2010).

138 Ibid.

139 Kommission Zukunft der Informationsinfrastruktur (2011).

In the EU GRDI 2020 report[140], it is assumed that over the next ten years, global research data infrastructures will have to be built in order to operate beyond linguistic, political and social boundaries. These infrastructures are to make research data available and support discovery, access and use. In this context, the model "Digital Ecosystems Science" will be introduced in which the following (new) stakeholders are involved: digital data libraries, digital data archives, digital research libraries and communities of research. This model implies a sometimes entirely new distribution of roles and tasks for the current stakeholders and calls for the creation of newly defined areas of responsibility. The main focus is always on the backup and re-use of research data, allowing the re-use of data also above and beyond disciplinary boundaries. To achieve these objectives, the report formulates eleven recommendations and courses of action which include among others the establishment of new professions and qualification processes. In addition, it is recommended to develop new tools (e.g. in the areas of data analysis or data visualization) and services (e.g. for data integration, for data retrieval or ontology services) for the management and use of data and to take "open science" and "open data" concepts into account.

The present comparative survey of the eleven academic disciplines in Germany confirms the validity of the above-mentioned statements. The overall picture reveals an urgent need for action, especially in the following areas:

- National and international programs have to be initiated to meet the new major challenges in the field of research data.
- A redefinition of roles and responsibilities is necessary to deal with the different areas of activity involved in the accessibility and re-use as well as the long-term preservation of research data.
- New careers and educational opportunities need to be developed and research data management has to be present in (new) degree programs and major study courses to ensure the professional handling of research data.

140 See GRDI 2020 Roadmap Report (2011).

- The publication of research data has to be regarded as an indispensable part of research processes in order to support the verification and further development of research results.

In conclusion, research data are both the result and the indispensable basis of scholarly work. They must be understood as resources that are continuously growing in importance for future generations of researchers as well as across disciplines. In this respect, they are part of the international cultural heritage. Therefore it is needed to curate and maintain research data throughout their entire life cycle.

Although there are already highly promising international approaches and some national developments and discussions have taken place in Germany, it will require a large, nationally coordinated effort before the vision of a "data infrastructure" can become reality. This process will involve both discipline-specific and interdisciplinary aspects and is to be embedded in international efforts. Legal, financial and organizational aspects should not hinder but support these developments. At this point, policy makers should get involved.